IJITKM Volume 8 • Number 1 June-Dec 2014 pp. 55-58 (ISSN 0973-4414)

Exploring Sentiment Classification Techniques in News Articles

Tirivangani BHT Magadza[1], Addlight Mukwazvure[2], K.P Supreethi[3]
Jawaharlal Nehru Technological University, Hyderabad, Telengana, India
bhtmagadza@gmail.com,addymkz@gmail.com,kppujari@gmail.com

Abstract: The emergence of web 2.0 applications has greatly contributed to the increase in volume of information available online today. User generated content can help organizations realize the demands of the public be it in e-commerce, politics or newsrooms. Sentiment analysis plays a pivotal role in the mining of such information thus it is a crucial tool not only in organizations' decision making process but also to the general users of a particular service. Most research on sentiment analysis focuses on subjective text like in micro-blogging, product and movie reviews. News articles sentiment analysis can be a bit difficult considering the need by journalists to remain neutral. Polarity of sentiments is not explicit therefore classification of people's sentiments in such a scenario is crucial. In this research we will outline the various methodologies used for polarity detection and analysis in news articles.

Keywords: Web 2.0 Applications, User Generated Content, Sentiment Analysis, Subjective Text, Classification, Polarity Detection.

1. Introduction

User opinions play an important role in the day to day operation of organizations. For effective service provision and decision making process, it is essential for organizations to have an idea of what users think regarding their products and services. Sentiment analysis is the tool that can be used to gather such important information.[1], defines sentiment analysis as "the computational study of people's opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics and their attributes." Once people's attitudes are known concerning a product or service in question, organizations can tailor make their services to meet people's expectations. Comments on news articles can be very valuable especially in politics. Understanding what readers think about a particular article can also help the newsagents modify their news content and coverage. Traditional methods of gathering reader opinions were mainly based on interviews and questioners. These however are time consuming and are sorely dependent on the good will of the reader [2]. With the emergence of web 2.0 applications however, it is now easier to gather news readers' comments from a global perspective and thus the need for automatic classification. The remainder of this article is organized as follows. Section 2 provides an overview of classification methods that have been used in news comments. The challenges and potential research directions are discussed in Section 3. Finally, we conclude and discuss future work in Section 4.

2. Literature Review

Research on sentiment analysis has been mostly centered on product/movie reviews as well as social network analysis. Analysis of news comments tend to be difficult considering that commentators can deviate from the topic in question as and also the fact that journalists tend to be neutral in their reporting[3] [4]. We will explore on the two commonly used methods for classification which are Machine Learning and Sentiment Lexicon approaches [2]. Machine learning methods are generally classified under supervised and unsupervised learning. Unsupervised learning methods do not use training data set for classification while classification using supervised learning algorithms involves two major phases which are a training phase and a test phase used for validation. There are various machine learning algorithms which include the K-Nearest Neighbor (kNN), Iterative Dichotomiser (ID3), Centroid classifier, Winnow classifier, Naive Bayes (NB), Support Vector Machines (SVM) and Maximum Entropy (MaxEnt).

2.1 General Overview of Machine learning algorithms

2.1.1 ID3

ID3 is a decision tree greedy algorithm developed by J. Ross Quinlan[5]. It has three parameters which are D (which is the data partition), attribute list, and attribute selection method. The data partition is initially, a complete set of training tuples and their associated class labels while the attribute list is a list of attributes which describe the tuples. The attribute selection method is used to specify a heuristic procedure for selecting the attribute that best discriminates the given tuples according to class [5].

2.1.2 Support Vector Machine

The Support Vector Machine is used for classifying both linear and nonlinear data. It uses a nonlinear mapping to transform the input data into a higher dimensional feature space and find an optimal hyperplane that maximizes the margin between the classes. SVM makes decisions based on the support vectors that are selected as the only effective elements in the training set[5],[6]. SVM can also be used for multi-class categorization but the major challenge with this is that it requires a voting scheme based on pair-wise classification results[6].

2.1.3 K Nearest Neighbor

kNN is based on comparing a given test tuple with training tuples that are similar to it. Each tuple represents a point in an *n*-dimensional space where n represents the attributes of each tuple. Given an unknown tuple (or text document), the classifier searches the pattern space for the *k* training tuples that are closest to the unknown tuple. These *k* training tuples are the *k* "nearest neighbors" of the unknown tuple. Closeness is based on some measure of distance which could be cosine or Euclidean distance based on the features of the tuples[5], [7]. The kNN classifier however tends to be labor intensive when presented with lager training sets[6].

2.1.4 Naive Bayes

Bayesian classification has its roots in statistical mathematics, thus its properties are mathematically provable[8]. It is based on Bayes theorem of posterior probability*: P (H/X*) where P represents the probability, H is the hypothesis and X is the evidence (which is the observed data tuple). The classifier supplies relative probabilities along with the class labels, which expresses the confidence of a decision. Given a training set and the associated class labels, each tuple will be represented by an *n* dimensional attribute vector, The classifier determines that the attribute vector belongs to a class with the highest posterior probability[5].

2.1.5 Maximum Entropy

Unlike the Naïve Bayes classification which assumes feature independence, Maximum Entropy classification does not assumes conditional independence of features. It is based on the Principle of maximum Entropy where from a number of computed models, the model with the maximum entropy over all models that satisfy the given constraints is selected for classification [9]. It is most appropriate for text categorization.

2.1.6 Centroid Based Technique

The k-means algorithm follows two basic approaches. The first step is the initialization of k cluster centroids. For the remaining objects, each object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster mean. The second step is to move and update the centroids by computing the new mean for each cluster. This process iterates until the criterion the function converges[5].

2.1.7 Winnow

[8], describes winnow as "an adaptive learning algorithm from equivalence queries, requiring the user to provide a counterexample to its current hypothesis." Normally used when many of the given features prove to be irrelevant. Winnow
uses linear threshold functions as hypotheses and performs incremental updates to its current hypothesis[10].

2.2 Classification Methods in News Articles

The classification methods used here are sentiment lexicon methods and machine learning approaches. Most of the above highlighted machine learning algorithms have been implemented in news classification with the most common ones being the SVM, Naïve Bayes, kNN and MaxEnt.

In [11], they adopted a machine learning approach to sentiment classification. The framework proposed first uses Ajax crawling technology to crawl comments from websites. After that, two classifiers (K-Nearest Neighbor and Support Vector Machine) were trained to classify online comments. The work majored on classification of Chinese news comments. Their corpus consisted of news and comments from *www.sina.com* and they majored on three topics which are: Top Stories (about spring festival transportation), Entertainment Events and Sports Coverage. Their results concluded that SVM performed better than kNN. Almost similar to this is an approach used in Sentiment Classification for Stock News. The financial data was crawled from Sina Finances and then preprocessed before a classifier was trained. Appraisal Theory was used for text sentiment classification[12]. The algorithms used are kNN, SVM and Naïve Bayes.

A machine learning approach to sentiment classification was also employed in the article entitled Sentiment Analysis of Online News using MALLET[13]. An approach almost similar to that used in [11] was used. The general data pre-processing was done and each text document in the corpus was represented by a word vector. Here their corpus was news articles from BBC and they centered on gay marriages. They divided the corpus into a training set and a testing set. MALLET was used to compare six classification algorithms which are Maximum Entropy classifier, Naïve Bayes classifier, a decision tree rule base, a decision tree with the C4.5 algorithm, the Winnow algorithm and the Balanced Winnow algorithm with the Naïve Bayes classifier giving the best results.

In [14], a semi-supervised approach to sentiment analysis of stock market news was used. Here, a polarity dictionary was used to determine polarity of the news. Bootstrapping was used to determine the article's polarity based on the frequency of words in the polarity dictionary [14].

Prashant Raina [15], described a sentic computing approach to sentiment analysis of news articles from MPQA corpus where he used ConceptNet and SenticNet common sense knowledge bases. Their idea was to implement sentence- level sentiment analysis. Their opinion engine comprised of a semantic parser, sentiment analyzer and a copy of SenticNet database. The semantic parser was used to extract common sense concepts from each sentence. The semantic analyzer then matched these concepts with sentic vectors in SenticNet. The sentic vector merely described the emotion in the sentence but not its polarity. A polarity measure was then used to convert a sentic vector into a polarity score between -1.0 and +1.0 [15]. The sentic vector of each concept's sentic vector was based on the Hourglass of Emotions [16] which groups emotions into four categories which are pleasantness, aptitude, attention and sensitivity [16], [17].

An unsupervised approach to sentiment analysis is generally based on a Sentiment Lexicon [18], [19]. A semantic orientation approach to text classification is presented in [19] where the proposed algorithm extracted phrases which contain adjectives or adverbs and then estimate the phrase's semantic orientation. The semantic orientation was then used to classify the reviews. Although in this case, this method was applied to product reviews, the same can be applied to news comments. To the best of our knowledge, the approach has not been used in the news domain prior to the writing of this article.

[20], used a semantic structure approach to target identification on online news text. They used FrameNet data to label an opinion holder. To alleviate FrameNet's limited number of words in the annotated corpus, they used a clustering technique to predict a most probable frame for an unseen word. It is however important to note the limitations posed by use of opinionated verbs and adjectives which include the fact that a sentence containing sentiment words may not express any sentiment[21]. Adjectives may indicate subjectivity but there may be insufficient context to determine semantic orientation thus if results are not suffer there is need to use a comprehensive sentiment lexicon[19]. Objectivity also needs to be considered when dealing with opinion lexicon otherwise some sentiments will be missed.

3. Challenges

3.1 Target Identification

Most research on news sentiment analysis focuses only on polarity detection without considering target extraction. It is easier to identify targets in product and movie reviews since users tend to focus on the product in question. In [22], a dependency parser in Chinese news, was used to identify opinion holders while [23] the centering theory was used. Both works focus on Chinese news texts and none bases on English texts.

3.2 Domain Specification

The major challenge with both supervised and unsupervised approaches to sentiment analysis is the aspect of domain dependence. A classifier trained for one domain normally performs poorly in another domain. Polarity detection using supervised methods is also time dependent[18].

4. Conclusions and Future Work

We have highlighted how both machine learning and semantic orientation are used for news classification. The news domain to sentiment analysis has less coverage so far. Most research has been centered on supervised learning methods. Considering the challenges highlighted in section 3, a hybrid implementation of sentiment analysis can curb those problems.[24], proposed a hybrid hierarchical classification in review texts. This hybrid approach can also be implemented in news articles sentiment analysis. We propose a simple approach to the hybrid implementation which combines Sentiment Lexicon and machine learning supervised methods to generate features which can then be used to train the classifier. Instead of ignoring opinion targets, algorithms like the Anaphora resolution technique can then be used for target identification.

References

[1] B. Liu and L. Zhang, "Chapter 13 - A Survey of Opinion Mining and Sentiment Analysis," in *Mining Text Data*, C. C. Aggarwal and C. Zhai, Eds. Springer US, 2012, pp. 415–463.

[2] E. Boiy, P. Hens, K. Deschacht, and M.-F. Moens, "Automatic Sentiment Analysis in On-line Text," in *Proceedings of the 11th International Conference on Electronic Publishing. Vienna, Austria*, 2007, no. June, pp. 349–350.

[3] A. Balahur, R. Steinberger, M. Kabadjov, V. Zavarella, E. Van Der Goot, M. Halkia, B. Pouliquen, and J. Belyaeva, "Sentiment analysis in the news," *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*,*Valletta, Malta,* pp. 2216–2220, 2010.

[4] M. C. Abhignya, S. Padmaja, S. S. Fatima, P. Kosala, and S. Bandu, "Comparing and Evaluating the Sentiment on Newspaper Articles□: A Preliminary Experiment," in *Science and Information Conference (SAI)*,*London*, 2014, pp. 789–792.

[5] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Second Edi. Morgan Kaufmann, 2006, p. 800.

[6] S. M. Namburu, H. Tu, J. Luo, and K. R. Pattipati, "Experiments on Supervised Learning Algorithms for Text Categorization," *Aerospace Conference*. IEEE, *Big Sky, MT*, pp. 1–8, 2004.

[7] C. C. Aggarwal, *Social Networks Data Analytics*. Springer, 2011, p. 502.

[8] A. Scime, *Web Mining: Applications and Techniques*. IGI Global, 2005, p. 442.

[9] E. Boiy, P. Hens, K. Deschacht, and M. Moens, "Automatic Sentiment Analysis in On-line Text," *ELPUB*, no. June, 2007.

[10] M. Balcan, "8803 Machine Learning Theory," 2011. [Online]. Available: http://www.cc.gatech.edu/~ninamf/ML11/. [Accessed: 12-Nov-2014].

[11] W. Fan and S. Sun, "Sentiment classification for online comments on Chinese news," in *Sentiment Classification for Online Comments on Chinese NewsComputer Application and System Modeling (ICCASM), 2010 International Conference on*, 2010, no. Iccasm, pp. 740–745.

[12] Y. Gao, L. Zhou, and Y. Zhang, "Sentiment classification for stock news," in *Pervasive Computing and Applications (ICPCA), 2010 5th International Conference on*, *Maribor,* 2010, pp. 99–104.

[13] S. Fong, Y. Zhuang, J. Li, and R. Khoury, "Sentiment Analysis of Online News Using MALLET," in *Computational and Business Intelligence (ISCBI), 2013 International Symposium on*, *New Dehli,* 2013, pp. 301–304.

[14] K. Mizumoto, H. Yanagimoto, and M. Yoshioka, "Sentiment Analysis of Stock Market News with Semi-supervised Learning," in *2012 IEEE/ACIS 11th International Conference on Computer and Information Science*, 2012, pp. 325–328.

[15] Prashant Raina and P. Raina, "Sentiment Analysis in News Articles Using Sentic Computing," *2013 IEEE 13th International Conference on Data Mining Workshops*,*Dallas,TX,* pp. 959–962, Dec. 2013.

[16] E. Cambria, A. Livingstone, and A. Hussain, "The hourglass of emotions," in *Cognitive behavioural systems*, vol. 7403 LNCS, Springer, 2012, pp. 144–157.

[17] R. Plutchik, "Nature of Emotions," *American Scientist*, vol. 89, pp. 344–350, 2001.

[18] B. Seyed-Ali, D. Andreas, S. S.-A. Bahrainian, and A. Dengel, "Sentiment Analysis Using Sentiment Features," in *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on*, *Atlanta,GA* 2013, vol. 3, pp. 26–29.

[19] P. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," *Proceedings of the 40th annual meeting on association for Computational Linguistics (ACL)*, no. July, pp. 417–424, 2002.

[20] S.-M. Kim, E. Hovy, M. Rey, and I. S. I. Edu, "Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text," in *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, 2006, no. July, pp. 1–8.

[21] B. Liu, "Sentiment Analysis and Opinion Mining," *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp. 1–167, May 2012.

[22] B. Lu, "Identifying opinion holders and targets with dependency parser in Chinese news texts," *Association for Computational Linguistics*, no. June, pp. 46–51, 2010.

[23] T. Ma, W. Xiaojun, and X. Wan, "Opinion Target Extraction in Chinese News Comments," in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, *Stroudsburg, PA,*2010, no. August, pp. 782–790.

[24] W. Wei and J. A. Gulla, "Sentiment analysis in a hybrid hierarchical classification process," in *Seventh International Conference on Digital Information Management (ICDIM 2012)*, *Valletta, Malta,* 2012, pp. 47–55.